Homework 11

36-467/667, Fall 2020

Due at 6 pm on Thursday, 19 November 2020

AGENDA: Working with compartment/population models, and Markov processes more broadly.

Now-common ideas like “early adopters” and “viral marketing” grew from sociological studies of the diffusion of innovations. One of the most influential of these studies tracked how a then-new antibiotic, tetracycline, spread among doctors in four small towns in Illinois in the 1950s. In this assignment, we will go back to that data to look at one of the crucial ideas, that of the innovation (prescribing tetracycline) “spreading” from person to person.

The study ran from November 1953, counted as month 1, to February 1955, counted as month 17. The data file `ckm_nodes.csv` records, for each doctor in those four towns, what town they lived in, in which month of the study period they started prescribing tetracycline, and a number of other variables about the doctor. If a doctor was already prescribing tetracycline when the study opened, or began prescribing it during November 1953, their adoption date was recorded as 1. If a doctor never prescribed tetracycline during the study, the adoption date is given as `Inf`. If it is unknown whether or not the doctor prescribed the drug during the study, the adoption date is `NA`.

1. (5) Clean the data set by removing all the doctors for whom the adoption data is unknown. Check that you have 125 doctors remaining.

2. (5) Create a new variable which records, for each month in the study, how many doctors *began* prescribing tetracycline in that month. Plot new adoptions over time. (Be careful — some doctors never started prescribing during the study.)

3. (5) Create a variable recording, for each month in the study, the number of doctors who were prescribing tetracycline by the end of that month.

4. (10) Create a variable recording, for each month in the study, the number of the doctors who *could* have begun prescribing during the next month (i.e., had not begun prescribing yet by the end of the month). Check that the sum of this variable and the number of already-prescribers is constant, and equal to the total number of doctors.

5. (5) Calculate, for each month $t \geq 2$, the fraction of not-yet-prescribing doctors who *did* begin prescribing. Plot this against t.

6. (5) For months $t \geq 2$, plot the fraction of new adopters (from Problem 5 in month t against the number of already-adopters (from Problem 3) in the previous month. Describe the shape of the scatter-plot.

7. (10) Let $I(t)$ be the number of doctors who have adopted tetracycline by month t, and $\hat{p}_{SI}(t+1)$ be the sample proportion of doctors who adopt in month $t+1$ (among doctors who had not yet adopted by the end of month t). Plot $\hat{p}_{SI}(t+1)$ against $I(t)$. Explain which mathematical variable corresponds to which variable in your code.

8. (5) In the SI model, $\hat{p}_{SI}(t)$ is, for each month, a proportion obtained from binomial trials. Estimate the variances in these proportions for each month $t \in 2:17$, and plot them over time. What is the ratio of the largest to the smallest variance that you estimate? *Hint:* The variance of a binomial with m trials and success rate p is $mp(1-p)$.

9. (9) Use `lm` to fit a model where $p_{SI} = \alpha I$, and report your estimate of α. *Hints:* Be careful that you don't fit a model of the form $p_{SI} = \alpha_0 + \alpha_1 I$, and use the previous problem to get weights.

10. (15) Write code which will simulate a model where each doctor who hasn't adopted yet in month t independently adopts or not in month $t+1$ with probability $\alpha I(t)$. The code should be an R function with inputs α, the total number of doctors n, the initial number of adopters $I(1)$, and the number of months to simulate T. It should return as output a vector of the number of doctors who have adopted by (not in) each month, comparable to what you calculated from the data in Problem 3. Check that your code gives the right results for the mean and variance of the number of new adoptions over multiple simulations. Your check should use small values of n and T, and a range of values of $I(1)$ and of α.

11. (10) Simulate the model from the previous problem 100 times, with n and $I(1)$ matched to the data, and α set to your estimate from Problem 9. For each month, find the 5 and 95 percentiles of $I(t)$ over the simulations. Plot these percentiles versus t, and add the actual value of $I(t)$ to the plot. What does the plot tell you about how well the model works?

12. (5) Repeat Problem 11, but double α. Why might the model work better at a different value of α than the one you estimated in Problem 9?

13. (1) How much time did you spend on this problem set?

PRESENTATION RUBRIC (10): The text is laid out cleanly, with clear divisions between problems and sub-problems. The writing itself is well-organized, free of grammatical and other mechanical errors, and easy to follow. Questions which ask for a plot or table are answered with both the figure itself and the command (or commands) use to make the plot. Plots are carefully labeled, with informative and legible titles, axis labels, and (if called for) sub-titles and legends; they are placed near the text of the corresponding problem. All quantitative and mathematical claims are supported by appropriate derivations, included in the text, or calculations in code. Numerical results are reported to appropriate precision. Code is properly integrated with a tool like R Markdown or knitr, and both the knitted file and the source file are submitted. The code is indented, commented, and uses meaningful names. All code is relevant, without dangling or useless commands. All parts of all problems are answered with coherent sentences, and raw computer code or output are only shown when explicitly asked for.

Extra credit I (5): A reasonable refinement of the model is that doctors who are in the same town are more likely to pass on information to each other than are doctors in different towns. Re-process the data to track adoptions separately in each town, and assume that the probability of adoption is proportional to the number of already-adopting doctors within the town (but with the same proportionality-factor α across towns). Report the new estimate of α, and simulate a model which tracks adoptions separately across the four towns. Does the new simulation seem to fit the data better?

Extra credit II (5): An even more refined model is that doctors learn about new information from specific other doctors whom they talk to. The data file `ckm_network` contains a matrix indicating which doctors were friends with each other. Can you re-estimate the model so that the probability of adoption depends on the fraction of a doctor's friends who have adopted? Do simulations of this model fit better?